HW2: Bias in Data and Prediction

DSCI 531 Spring 2024

Due at 4pm PT Feb 7th, 2024

1 Overview

Many of the current databases are influenced by the stereotypes in people's perceptions, during data selection and labeling. For example, the data used to train a resume-filtering system might be biased towards a specific gender and the underlying model might capture such bias when being trained on such data. However, gender should not play a role in an unbiased resume-filtering system. The first step toward this goal is to simply become aware of any existing biases using statistical analysis and try some simple ways to remove such bias.

In this exercise, we investigate possible bias in data and how such bias will affect the model's decisions. We also investigate several simple ways to avoid unintended biases for achieving a fairer model. Bias in the data is one of the sources of bias that can affect the fairness of a machine learning model.

2 Dataset

We will experiment with tabular data and use two datasets, namely Adult[1] and German Credit[2]. Both datasets are very commonly used in the field of AI fairness.

- Adult is census data. The goal is to predict whether the individual's income exceeds 50K/yr based on other features (1 → income > 50K/yr, 0 → income <= 50K/yr). The protected feature is gender (0 → male, 1 → female).
- German dataset classifies people described by a set of attributes as good or bad credit risks (1 → good credit, 0 → bad credit). The protected feature is age (1 → age >= 33, 0 → age < 33).

Please visit the webpages for more details about the datasets.

3 Tasks

1. Understand the definitions of statistical parity and equalized opportunity [1]. Implement the two metrics with Python. Report the results on the given test cases with your implemented functions.
2. Preprocess the datasets and convert the features to numbers so they can be handled by algorithms. You should handle categorical features and numerical features differently.
3. Understand the bias in data. Investigate whether the protected feature is correlated with the outcome. 1) Compare the means of the outcome across different protected groups. 2) Perform a t-test on the outcome between the two protected groups. Report the p-value. Is the test result significant?
4. Understand the bias in prediction. The bias in data will be propagated to the model if the model is trained on such data. Train an ML model to predict the outcome from the features on the training data. Report the accuracy and the two fairness metrics on the test data.

[1] https://archive.ics.uci.edu/ml/datasets/adult

[2] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

5. Explore several naive ways to mitigate bias in prediction. 1) Because we want the model's predictions to be independent of the protected feature, one simple way is to just exclude the protected feature when training the model. Try this and report the accuracy and two fairness metrics again. How are the results different from Step 4? Are the predictions fairer now? Does it make the accuracy drop a lot? How do you explain it? 2) Another straightforward method to counteract the effect of the protected feature in model predictions is data augmentation [2]. Specifically, for every data sample, we create a new sample having the same features (except the protected attribute(s)) and label as the original sample but with the opposite protected attribute value. An example is shown in Figure 1 of [2]. The synthetic data is used to augment the original training data. The concatenated data (synthetic data + original training data) is used to train the model. Note that in this process the test data is not touched. Implement this idea and redo Step 4 and report the results. According to your results, is this method effective in mitigating bias? How does it affect the accuracy? How do you interpret the results?

You will be provided with a Jupyter nobebook with detailed instructions for each task. The cleaned datasets will also be provided. Please do not use datasets from other sources. Complete the TODOs in the notebook. The notebook and the data can be downloaded here. You are not allowed to use the external fairness-oriented packages, such as ai360. Please include as many comments about your code as possible. You should run every cell and keep the outputs before submitting the notebook.

4 Helpful Resource

- A Tutorial on Fairness in Machine Learning

5 Submission Guideline

Please submit your notebook file named as hw2-lastname-firstname.ipynb to Blackboard before 4pm Feb 7th, 2024.

References

[1] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR) 54(6), 1–35 (2021)

[2] Sharma, S., Zhang, Y., R´ıos Aliaga, J.M., Bouneffouf, D., Muthusamy, V., Varshney, K.R.: Data augmentation for discrimination prevention and bias disambiguation. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. pp. 358–364 (2020)